Research Article

Challenges Integrating Skin Sensitization Data for Assessment of Difficult to Test Substances

Allison Greminger¹, Joseph Frasca², Katya Goyak¹ and Colin North¹
¹ExxonMobil Biomedical Sciences, Inc., Health and Environmental Sciences Division, Annandale, NJ, USA; ²ExxonMobil Biomedical Sciences, Inc., Health and Environmental Sciences Division, Spring, TX, USA

Abstract

Difficult to test substances including poorly soluble, mildly irritating, or Unknown or Variable Composition Complex reaction products or Biological Materials (UVCBs), producing weak or borderline in vivo results, face additional challenges in in vitro assays that often necessitates data integration in a weight of evidence (WOE) approach to inform skin sensitization potential. Here we present several case studies on difficult to test substances and highlight the utility of Toxological Prioritization Index (ToxPi) as a data visualization tool to compare skin sensitization biological activity. The case study test substances represent two poorly soluble substances, tetrakis (2-ethylbutyl) orthosilicate and decyl palmitate, and two UVCB substances, alkylated anisole and hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2 undecanone. Data from key events within the skin sensitization adverse outcome pathway were gathered from publicly available sources or specifically generated. Incorporating the data for these case study test substances as well as on chemicals of a known sensitization class (sensitizer, irritating non-sensitizer, and non-sensitizer) into ToxPi produced biological activity profiles which were grouped using unsupervised hierarchical clustering. Three of the case study test substances concluded to lack skin sensitization potential by traditional WOE produced biological activity profiles most consistent with non-sensitizing substances, whereas the prediction was less definitive for a substance considered positive by traditional WOE. Visualizing the data using bioactivity profiles can provide further support for WOE conclusions in certain circumstances but is unlikely to replace WOE as a stand-alone prediction due to limitations of the method including the impact of missing data points.

1 Introduction

Assessment of skin sensitization potential is a required hazard endpoint in many chemical registration systems globally. It can also be an important parameter in research and development, as the presence or absence of skin sensitization potential can dictate whether a chemical is suitable in different end uses, and consequently whether investment in development is warranted. Hence, correct assessment of chemicals for sensitization has important implications both for product safety/stewardship, and investment decisions for new chemical development. Historically skin sensitization testing has relied on laboratory animal models measuring elicitation, such as the Guinea Pig Maximization Test (Magnusson and Kligman, 1969) and the Buehler Test (Buehler, 1965), or those measuring induction, such as the Local Lymph Node Assay (LLNA) (Kimber et al., 1994). In some cases skin sensitization testing has been performed in human volunteers via Human Repeat Insult Patch Testing (McNamee et al., 2008). Over time the LLNA became the predominant test used to assess skin sensitization potential, as it provided multiple advantages over other methods, such as increased sensitivity and specificity, quantitative measures of potency, and reduced animal use (Gilmour et al., 2019; Kimber et al., 1995).

Even with these advantages, there are still occasions where it is difficult to conclude on sensitization potential from LLNA data due to uncertainties around potential false positives, predictive performance vs observed human and guinea pig responses, and technical limitations such as solvent-dependent responses (Anderson et al., 2011). For instance, the LLNA has been recognized both for its limited ability to distinguish irritants and weak sensitizers (Adenuga et al., 2012; Basketter et al., 1998; Montelius et al., 1998) and for its inability to consistently identify chemicals producing borderline or weak results in which the “true” biological activity is close to the classification threshold (Kolle et al., 2013). Classification threshold here is defined relative to the respective regulatory classification scheme such as the Globally Harmonized System of Classification and Labeling of Chemicals (UN GHS)
or in the European Union regulation (EC) no 1272/2008 on classification, labelling, and packaging (EU CLP) of substances and mixtures. In these borderline instances, an individual LLNA result may fall above or below the classification threshold, potentially leading to discordant conclusions regarding the likelihood to produce a sensitization response (Dimitrov et al., 2016a; Hoffmann, 2015; Kolle et al., 2013). Recent evaluation of the predictive performance of the LLNA compared to human allergic reactions has reported a balanced accuracy of 58%, and none of the strong skin sensitizers in the human reference data were incorrectly predicted by the LLNA (OECD, 2021). It is foreseen that additional mechanistic information acquired from New Approach Methodologies (NAM) can be incorporated to increase confidence in characterization of skin sensitization.

Significant progress has been made to develop predictive computational and in vitro models based on well-characterized mechanisms of skin sensitization. As a result, there are widely available Organization for Economic Co-operative Development (OECD) guideline tests which address specific key events (KE) in the skin sensitization adverse outcome pathway (AOP) and, when used in combination, appear to equal the predictive performance of in vivo laboratory animal tests (Strickland et al., 2016). Briefly, the AOP consists of four KEs and captures a simplistic description of the cascade of biological events responsible for chemical mediated skin sensitization characterized by protein binding, cell activation, and subsequent elicitation of an immune reaction (Figure 1). For example, the Direct Peptide Reactivity Assay (DPRA) addresses the molecular initiating event in the sensitization pathway, the covalent binding of a hapten to skin proteins (Gerberick, 2016; OECD, 2023a). At the cellular level, KE2 is the activation of a keratinocyte response, which can be assessed in the KeratinoSens™ or LuSens™ assays (Natsch et al., 2011; Ramirez et al., 2014; OECD, 2022); while KE3, the activation of dendritic cell response, can be assessed in the human Cell Line Activation Test (h-CLAT), interleukin 8 luciferase (IL-8 LUC), U-SENS™ (Ashikaga et al., 2006; OECD, 2023b) or GARD™skin (EURL ECVAM Scientific Advisory Committee, 2021) (Figure 1). Although the GARD™skin assay has been incorporated into the OECD TG aligned with KE3, it should be noted that the gene signature profile used in the decision context of this assay includes genes associated with KEs across the skin sensitization AOP (Roberts, 2018). Finally, in-silico screening tools, such as OASIS TIMES or QSAR Toolbox, supplement our understanding of chemicals of interest by identifying structural fragments associated with known skin sensitization mechanisms in parent structures as well as probable metabolites occurring in skin or by autoxidation (Dimitrov et al., 2016b; Patlewicz et al., 2014; Ivanova et al., 2020).

Despite the mechanistic understanding of the biological pathway and associated tests, these KE-based tests each have clearly defined requirements for test substance properties that impact feasibility or confound reliance on negative results particularly for difficult to test substances (see Table 1 for a non-exhaustive list). In our experience, difficult to test substances include UVCBs and/or those that have challenging physicochemical properties (e.g., low water solubility, lipophilic). For example, evidence indicates that the h-CLAT assay demonstrates low predictive performance for lipophilic compounds with a high octanol-water partition coefficient (LogKow > 3.5) compared to the LLNA increasing the likelihood of a false negative result (Takenouchi et al., 2013); while the DPRA is not applicable to UVCBs due to the need for an exact ratio of test substance to peptide which is made...
Tab. 1: Advantages and limitations for difficult to test substances of currently available OECD TG skin sensitization test methods

<table>
<thead>
<tr>
<th>Key Event</th>
<th>Relevant Test Methods</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covalent Protein Binding</td>
<td>(Q)SAR/StructuralAlerts</td>
<td>Non-test method</td>
<td>Limited applicability domain Requires defined molecular structure(s)</td>
</tr>
<tr>
<td></td>
<td>Electrophilic Chemistry</td>
<td>Non-test method</td>
<td>Limited predictivity</td>
</tr>
<tr>
<td></td>
<td>OECD442C: DPRA</td>
<td>Smaller requirement for test material compared to in-vivo</td>
<td>Requires defined molecular weight Limited by water solubility Limited choice of solvents</td>
</tr>
<tr>
<td></td>
<td>OECD TG 442C: ADRA</td>
<td>Smaller requirement for test material compared to DPRA Can be conducted via gravimetric approach without defined molecular weight</td>
<td>Limited choice of solvents</td>
</tr>
<tr>
<td></td>
<td>OECD TG 442C: kDPRA</td>
<td>Smaller requirement for test material compared to in-vivo</td>
<td>Limited choice of solvents</td>
</tr>
<tr>
<td>Keratinocyte Activation</td>
<td>OECD TG 442D: KeratinoSens™/LuSens</td>
<td>Smaller requirement for test material compared to in-vivo</td>
<td>Limited choice of solvents</td>
</tr>
<tr>
<td>Dendritic Cell Activation</td>
<td>OECD TG 442E: h-CLAT</td>
<td>Smaller requirement for test material compared to in-vivo</td>
<td>Limited choice of solvents Log Kow &gt; 3.5 confounds negative results</td>
</tr>
<tr>
<td></td>
<td>OECD TG 442E: U-SENS™</td>
<td>Smaller requirement for test material compared to in-vivo</td>
<td>Limited choice of solvents</td>
</tr>
<tr>
<td></td>
<td>OECD TG 442E: GARD™skin</td>
<td>Expanded selection of solvents Available reference library of difficult to test substances</td>
<td>Regulatory interpretation limited to Dendritic Cell Activation</td>
</tr>
<tr>
<td>T-Cell Proliferation</td>
<td>OECD TG429: LLNA</td>
<td>Difficult to test substances can be applied neat</td>
<td>False positives due to irritation or specific chemistry</td>
</tr>
<tr>
<td></td>
<td>OECD TG 442A: LLNA-DA</td>
<td>Difficult to test substances can be applied neat</td>
<td>False positives due to irritation or specific chemistry</td>
</tr>
<tr>
<td></td>
<td>OECD TG 442B: LLNA-BrdU</td>
<td>Non-radioactive</td>
<td>No regulatory threshold for hazard classification subcategorization (OECD 442B)</td>
</tr>
<tr>
<td>Elicitation</td>
<td>OECD TG 406: Buehler Assay/ GPMT</td>
<td>Measures adverse outcome</td>
<td>Animal/time intensive</td>
</tr>
</tbody>
</table>

difficult by the unknown or variable molecular weights of UVCB substances (EURL-ECVAM, 2012). While recent development of a gravimetric based test method for detecting covalent binding with proteins (amino acid derivative reactivity assay) may allow testing of test chemicals without a known molecular weight, however the approved solvents are the same as the traditional DPRA likely leading to solubility concerns. Additionally, quantitative structure activity relationship (QSAR) models typically have applicability domains that do not include UVCBs or substances on the extremes of physicochemical property boundaries (OECD, 2021). Recognizing the uncertainties associated with relying on a single assay to determine skin sensitization potential, three defined approaches (DA) that represent transparent, rule-based systems to data interpretation of KE-based NAMs have been published by the OECD and have been shown to better predict human skin sensitization hazard and potency when compared to the LLNA alone (Hoffmann et al., 2018; OECD, 2021; Kleinsteuer et al., 2018). Nonetheless, application of the two published DAs remain challenged for difficult to test substances due to the specific test method requirements described above as well as a higher burden of proof for negative results in certain circumstances. Therefore, difficult to test substances which produce borderline or weak results in the LLNA face additional challenges in KE-based in vitro assays that ultimately necessitates the application of WOE approaches to conclude on skin sensitization potential.

In this study, we present several real-world experiences in assessing challenging substances both in an overall WOE as well as utilizing the ToxPi framework as an additional tool to compare biological activity as it relates to skin sensitization. The ToxPi graphical user interface enables integration of multiple data sources and types (e.g. chemical/physical properties, test data, exposure information) by transforming data into visual rankings (Marvel et al., 2018). An advantage of this approach is its flexibility in assessing various parameters of biological activity and has been used previously to assist grouping of complex substances (Grimm et al., 2016, 2019), as well as ranking and prioritizing of chemicals for further study (Reif, et al. 2010, Tilley, et al. 2017). While this approach facilitated the comparison of several challenging test chemicals to prototypical sensitizers and non-sensitizers in a comprehensive assessment of biological activity, several limitations were identified in its use for supporting hazard classification decisions.
2 Materials & Methods

2.1 Test Chemicals
Test samples were obtained as follows: decyl palmitate (CASRN 42232-27-9), alkylated anisole (a reaction mass resulting in anisole attached to one or more C8-C10 branched hydrocarbon chains, no relevant CASRN) and hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2 undecanone (CASRN 68411-85-8) from ExxonMobil Chemical Company (Houston, TX USA); tetakis (2-ethylbutyl) orthosilicate (CASRN 78-13-7) from Gelest, Inc. (for LLNA, h-CLAT and DPRA) or from Sigma Aldrich (for GARD™skin); p-phenylenediamine (PPD) (CASRN 106-50-3), nonanoic acid (CASRN 112-05-0) and methyl salicylate (CASRN 119-36-8) from Sigma Aldrich for GARD™skin.

The following data were obtained from SkinSensDB1, a curated database of skin sensitization assay data (Wang et al., 2017); specifically, data from OECD TG 442C- DPRA, OECD TG 442D- KeratinoSens™/LuSens™, OECD TG 442E- h-CLAT and OECD TG 429/442B-LLNA were incorporated for p-phenylenediamine, aniline (CASRN 62-53-4), linalool (CASRN 78-70-6), eugenol (CASRN 97-53-0), 2-mercaptopbenothiazole (2-MBT) (CASRN 149-30-4), and cinnaamal (CASRN 67485-29-4) as prototypical skin sensitizers; p-aminobenzoic acid (CASRN 150-13-0), methyl salicylate, propylene glycol (CASRN 57-55-6), isopropanol (CASRN 95-47-6), salicylic acid (CASRN 69-72-7), and dextran (CASRN 9004-54-0) as prototypical non-sensitizers; and lactic acid (CASRN 50-91-5), hexane (CASRN 110-54-3), nonanoic acid, sodium lauryl sulfate (CASRN 151-21-3), and xylene (CASRN 1330-20-7) as irritating non-sensitizers. GARD™skin assay results for all prototypical comparison substances were obtained from published sources (Johansson et al., 2014, 2017).

2.2 Local Lymph Node Assay (LLNA)
The LLNA BrdU-ELISA for tetrakis (2-ethylbutyl) orthosilicate and hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2 undecanone was conducted at MB Research (Spinnerstown, PA USA) under GLP standards. The study protocol was reviewed by the Institutional Care and Use Committee of MB Research and found compliant with acceptable ethical standards of animal welfare. The study was designed and conducted in accordance with OECD TG 442B and the Guide for Care and Use of Laboratory Animals: Eighth Edition, National Academies Press (OECD, 2018). Female CBA/J mice (31-41 total) were in the weight range of 18.5 to 23 g and were 8 to 12 weeks old at study initiation; and randomly assigned to dose group upon receipt using a non-selective allocation to equalize variation in body weight. Prior to the start of treatment (5 days acclimatization), animals were examined to ensure there were no observable skin lesions. Based on initial irritation screening in a single mouse to determine concentrations which would not lead to systemic toxicity or excessive local irritation, decyl palmitate was tested at 0%, 10%, 25% and 50% (v/v) in butanone in five female CBA/Ca mice per dose group; alkylated anisole was tested at 0%, 25%, and 50% (v/v) in acetonone:olive oil (4:1, AOO) and undiluted (100%) in five female CBA/Ca mice per dose group. Five days following the first topical application of test item, all mice were injected via tail vein with 20 μCi 3H-methyl thymidine (3HTdR). Animals were sacrificed five hours following administration of 3HTdR, and the draining auricular lymph nodes were excised and combined for each animal into a single cell suspension. 3HTdR incorporation was measured by scintillation counting reported as disintegrations per minute (DPM). The Stimulation Index (SI) was calculated by dividing the mean DPM of each animal by the mean DPM of the vehicle control group. The SI was regarded as a sensitizer if any treatment group yielded an SI value equal to or greater than 3 according to OECD TG 429.

The LLNA BrdU-ELISA for tetrakis (2-ethylbutyl) orthosilicate and hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2 undecanone was conducted at MB Research (Spinnerstown, PA USA) under GLP standards. The study protocol was reviewed by the Institutional Care and Use Committee of MB Research and found compliant with acceptable ethical standards of animal welfare. The study was designed and conducted in accordance with OECD TG 442B and the Guide for Care and Use of Laboratory Animals: Eighth Edition, National Academies Press (OECD, 2018). Female CBA/J mice (31-41 total) were in the weight range of 18.5 to 23 g and were 8 to 12 weeks old at study initiation; and randomly assigned (according to simple randomization) to dose group upon receipt using a non-selective allocation to equalize variation in body weight. Prior to the start of treatment (5 days acclimatization), animals were examined to ensure there were no observable skin lesions. Based on initial irritation screening in 2 mice per dose to determine concentrations which would not lead to systemic toxicity or excessive local irritation, tetrakis (2-ethylbutyl) orthosilicate was tested at 0%, 10%, 25% and 50% (v/v) in acetonone:olive oil (4:1, AOO) in five female CBA/J mice per dose group and hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2 undecanone was tested at 0%, 10%, 25% 50% and 100% (v/v) in acetonone:olive oil (4:1, AOO) in five female CBA/J mice per dose group. Five days following the first topical application of test item, all mice were administered bromodeoxyuridine (BrdU) by intraperitoneal injection. Animals were sacrificed 24 hours following administration of BrdU, and the draining auricular lymph nodes were excised and combined for each animal into a single cell suspension. BrdU incorporation was measured using a BrdU-specific Enzyme-Linked Immunosorbent Assay (ELISA) kit and reported as optical density values. The SI was calculated by dividing the mean optical density of each animal by the mean optical density of the vehicle control group for test substance concentrations less than 100% or a sham control group for 100% test substance concentration. The mean SI and standard deviation (S.D.) were calculated for each group from the individual animal data. A test article was regarded as a sensitizer if any treatment group yielded an SI value equal to or greater than 1.6 according to OECD TG 442B. Both OECD TG have undergone extensive inter-laboratory validation and have been shown to reliably detect moderate to strong sensitizers, thus no additional efforts to minimize potential confounders were applied to the study design. No blinding of group allocations was required for these studies as the data is determined by digital measurement and is not required by the OECD TG.

1 http://cwtung.kmu.edu.tw/skinsensdb
2.3 GARD™ skin
GARD™ skin studies were facilitated by SenzaGen AB (Lund, Sweden) from start to finish for all case study chemicals except in the case of Tetrakis-(2-ethylbutyl) orthosilicate where exposure of the SenzaCells and RNA extraction was performed by MB Research Labs (Spinnerstown, PA U.S.A) and then sent to SenzaGen for genomic analysis and data interpretation. All protocols associated with the GARD™ skin assay have been previously published and were aligned with the recently adopted protocol incorporated into OECD TG 442E (Forreyrd et al., 2016; Johansson et al., 2013; OECD, 2023b). In short, cultivated human dendritic cell-like SenzaCells (ATCC Patent Deposit Designation PTA-123875, obtainable exclusively from SenzaGen), are exposed in vitro to the test substances of interest for 24 hours. Following concentration response measurements of induced cell toxicity, an appropriate and test chemical specific input concentration was defined as the concentration inducing relative viability closest to 90%. Tetrakis-(2-ethylbutyl) orthosilicate and alkylated anisole were adequately solubilized in ethanol and tested at as tested at 500 µM. Hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2 undecanone, decyl palmitate, methyl salicylate, and nonanoic acid were adequately solubilized in DMSO and tested at 50 µM.

Transcriptional levels of the GARD™ skin gene prediction signature (Johansson et al., 2011) were quantified using the Nanostring nCounter system. The data are analyzed by a Support Vector Machine (SVM) (Cortes and Vapnik, 1995), which has been appropriately trained on samples generated during technology platform transfer, consisting of expression profiles generated by the reference panel of chemicals as described in Forreyrd et al. (2016). Final prediction calls are derived from the mean SVM decision value (Johansson et al., 2019) generated by the biological triplicate samples; any test chemical inducing a mean SVM decision value ≥0.5 is classified as a skin sensitizer. Consequently, any test chemical inducing a mean SVM decision value <0 is classified as a non-sensitizer.

2.4 Direct Peptide Reactivity Assay (DPRA)
The DPRA was conducted on tetrakis (2-ethylbutyl) orthosilicate by EAG Laboratories (Easton, MD, USA) under GLP standards according to OECD TG 442C (OECD, 2023a). Three replicate solutions of the test substance were incubated with cysteine or lysine containing peptides at a molar ratio of 1:10 or 1:50, respectively. Relative peptide concentration was measured by high performance liquid chromatography (HPLC) with gradient elution and ultraviolet (UV) detection at 220 nm following 24 ± 2 hours of incubation with the test chemical at a temperature of 25 ± 2.5°C. The concentration of the cysteine and lysine peptides were determined in the appropriate reference control samples by substituting into the peak area responses into the appropriate linear regression equation. The percent peptide depletion was determined in each test substance assay and positive control sample by measuring the peak area of the appropriate peptide and dividing it by the mean peak area of the relevant reference control samples.

2.5 Human Cell Line Activation test (h-CLAT)
The h-CLAT for tetrakis (2-ethylbutyl) orthosilicate and hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2 undecanone was conducted at MB Research Labs (Spinnerstown, PA USA) under GLP standards according to OECD Test Guideline 442E except where noted below (OECD, 2023b). Activation of a human monocytic cell line (THP-1; ATCC No. TIB-202) was evaluated by measuring changes in the expression of dendritic cell surface markers (CD86, CD54) via flow cytometry. For tetrakis (2-ethylbutyl) Orthosilicate, the test guideline was modified to utilize 100% Ethanol (1% final dilution) as the vehicle solvent since initial solubility studies indicated the test substance was not soluble in either DMSO or saline (0.9%). Additional solvent controls (1% Ethanol in RPMI-10 culture medium), as well as proficiency controls diluted in ethanol (isopropanol CARN: 67-63-0 non-sensitizer; 2-mercaptoethanol CASRN: 149-30-4 moderate sensitizer) were tested concurrently with the negative control (Lactic Acid; CASRN 50-21-5, ≥ 85% purity) and both positive controls (1-chloro-2,4-dinitrobenzene DNCB CASRN 97-00-7, ≥ 99% purity and Nickel Sulfate CASRN 10101-97-0, ≥ 99% purity) in the study to confirm no interference of the alternative vehicle. For hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2 undecanone no modifications to the test guideline were implemented as the test substance was soluble up to 500 mg/mL in DMSO.

THP-1 cells were incubated with the test substance in concentrations up to 5,000 µg/mL. After 24±0.5 hours of exposure, CD54/86 expression was measured by FITC-labelled anti-CD86, anti-CD54 or anti-mouse IgG1 (isotype control) antibodies at 4°C for 30 min. Cell viability was measured by propidium iodide staining. Results were quantified as the effective concentration inducing a relative fluorescence intensity (RFI) above the positive threshold criteria for each cell surface marker (RFI = 150 for CD86, and RFI = 200 for CD54). The substance was predicted as positive if an RFI greater than either threshold was observed for any test dose (with cell viability > 50%) in at least two of three independent assays, otherwise it was considered negative.

2.6 KeratinoSens™
The KeratinoSens™ assay for hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2 undecanone was conducted at Cyprotex (Watertown, MA USA) according to OECD TG 442D under non-GLP conditions as no laboratories had GLP-compliant capacity for the TG at the time of testing (ca. 2017) (OECD, 2022). Activation of KeratinoSens™ cells was evaluated by measuring the induction of luciferase gene expression of the Keap1-Nrf2-ARE pathway. KeratinoSens™ cells were incubated with the test substance for 48 hours in concentrations up to 500 µg/mL in DMSO, lysed and assessed for luciferase reporter gene expression using a luminescent assay (Luciferase Assay Kit, Promega). Cell viability was assessed using 3-(4,5-Dimethylthiazol-2-yl)-2,5-Diphenyldtetrazolium Bromide (MTT). Cinnamic aldehyde (CASRN 14371-10-9) and sodium lauryl sulfate were utilized as a positive and negative control respectively. The substance was predicted as positive if the reporter gene expression was induced above a fold of 1.5 at concentrations below 1000 µM (with cell viability > 70%) in at least two of three independent assays, otherwise it was considered negative.
2.7 \textit{In-silico}

OECD QSAR Toolbox 4.4\textsuperscript{2} was utilized to check for skin sensitization structural alerts (SA) of all chemicals presented here as well as metabolites by autoxidation or skin metabolism. A QSAR prediction for skin sensitization was generated using the Skin Sensitization with autoxidation v. 20.24 model of the OASIS TIMES v.2.29.1 software. Values for the energy of the Highest Occupied Molecular Orbital (eHOMO, eV) and Softness (σ) were obtained from OASIS TIMES and included in the ToxPi data integration under KE1 as these molecular parameters have been shown to inform nucleophile/electrophile interactions as they relate to skin sensitization potential (LoPachin et al., 2012, 2019; Mekenyan et al., 1997; Ouyang et al., 2014).

2.8 Data integration in Toxicological Priority Index (ToxPi)

The data were grouped based on their associated key event within the skin sensitization AOP; the framework for profiling and data visualization is detailed in Figure 2. ToxPi is represented as component slices of a unit circle with each slice representing a piece of information. For each slice, the distance from the origin is proportional to a normalized value, and the width represents the relative weight of that piece of information within the overall ToxPi profile. In the implementation presented here, all KE slices were weighted equally and normalized so that larger pie slices correlate to increased biological activity as it relates specifically to skin sensitization. For example, in Figure 2, where the orange slice represents LLNA results, a smaller EC 3 value (higher potency) correlates to a larger pie slice (further distance from origin). The pie slice representing physicochemical properties was weighted at half of the KE slices since the correlation to biological activity was considered less definitive. Raw data was log transformed and normalized using the min and max values for each endpoint; missing data points were entered as “NA” and not included in the dataset. Low-confidence results (either negative or positive) were included in the ToxPi. See supplemental information\textsuperscript{3} for raw data, normalization and scaling of each endpoint. Hierarchical clustering of ToxPi profiles was organized using the complete linkage model. The clustering methods in ToxPi GUI correspond to unsupervised methods available in the R function hclust using Euclidean distance (Marvel et al., 2018).

\textsuperscript{2} www.qsartoolbox.org
\textsuperscript{3} doi:10.14573/altex.2201122s
3 Results

3.1 WOE for Tetrakis (2-ethylbutyl) Orthosilicate

Problem formulation
The skin sensitization potential of tetrakis (2-ethylbutyl) orthosilicate was assessed to determine the appropriate hazard classification for regulatory registration purposes. Because the main objective was regulatory registration, the testing strategy was designed to meet the regulatory information requirements, following the decision logic recommended by the competent authority guidance for skin sensitization hazard identification (ECHA, 2017a; United Nations, 2019). Annex VII of the Registration, Evaluation, and Authorization of Chemicals (REACH) regulation stipulates that in vitro assays should be considered prior to conducting animal-based assays (European Parliament Regulation EC No 1907/2006).

NAM data generation
No structural alerts for sensitization potential were identified by the OECD QSAR Toolbox. TIMES Skin Sensitization with autoxidation model predicted a negative skin sensitization result (see supplemental data), however this result is considered as low confidence since the substance is outside of the model domain due to the presence of 30% unknown structural fragments.

Initial data generation focused on the mechanistically based in chemico and in vitro methods associated with key events 1 and 3 in the skin sensitization AOP (Figure 1); namely, a DPRA to assess covalent interaction with cysteine and lysine peptides in skin and the h-CLAT to assess the activation of dendritic cells through measurement of markers CD86 and CD54. In the DPRA, a precipitate was observed immediately upon addition of tetrakis (2-ethylbutyl) orthosilicate to the peptide solution, due to low aqueous solubility of the test chemical. Although the mean percent depletion values for cysteine and lysine (1.12 ± 1.0%; N=3, CV=91.7%) and 0.251 ± 0.43% (N=3, CV = 173%), respectively) predict no or minimal reactivity with skin proteins, the observed precipitate decreases confidence in this prediction, as the amount of chemical available in solution to react with the peptides is unknown. As per OECD Test Guideline 442C, no firm conclusion on the lack of reactivity should be drawn from a negative result in this instance. Therefore, the result was considered negative with low confidence for the purposes of WOE. In the h-CLAT, tetrakis (2-ethylbutyl) orthosilicate was tested in 100% ethanol, due to insolubility in either of the OECD TG validated solvents (saline and DMSO). In this modified version, the EC150 and EC200 for tetrakis (2-ethylbutyl) orthosilicate was >5000 μL/mL (highest dose tested). However, OECD TG 442E notes that substances with high octanol water partition coefficients (Log Kow > 3.5) tend to produce false negative results (Takenouchi et al., 2013). Based on a measured partition coefficient (Log Kow) >6.01, the result was considered negative with low confidence. Further information on dendritic cell activation (KE3) was generated using the GARD®skin assay which predicted tetrakis (2-ethylbutyl) orthosilicate (in ethanol) to be a non-sensitizer based on a mean SVM decision value of -1.07 (Table 2); all three replicates produced SVM decision values below zero, the established threshold for a positive result.

Animal data generation
Because the WOE based on QSAR and in vitro assays was considered negative but low confidence for classification purposes, tetrakis (2-ethylbutyl) orthosilicate was tested in an LLNA BrdU-ELISA (OECD TG 442B). The test substance was applied to CBA/J mice at 10%, 25%, and 50%, which produced stimulation indices (SI) of 1.1, 2.5, and 4.9, respectively (Tab. 2). However, it should be noted that the 50% concentration produced a mean ear thickness of >25% indicating “excessive local irritation” according to the OECD TG 442B. Although the substance would typically be classified as a skin sensitizer based on the results of the OECD TG criteria, it is evident that a positive dermal irritation response occurred in the LLNA BrdU-ELISA. It should be considered that the SI value of 4.9 produced by the 50% concentration level may be at least partially influenced by irritation. Considering an SI=1.6 as a positive indication of skin sensitization in this assay, the effective concentration (EC) at which the test substance may produce skin sensitization (i.e.EC 1.6) is calculated to be 15.5%. While GHS criteria does not yet exist for subcategory into high potency or low/moderate potency based on an SI of 1.6, the magnitude of the dose response is suggestive of low to moderate potency.

Evidence integration
Evidence for tetrakis (2-ethylbutyl) orthosilicate was inconclusive for protein binding (KE1) based on QSAR and DPRA due to technical limitations with this specific chemistry (out of domain for QSAR; negative but low confidence results in DPRA due to precipitation). Data was not generated for KE2 as it would likely not have provided additional conclusive information given the physicochemical challenges of the substance. Overall, the test substance was considered negative for dendritic cell activation (KE3) based on an inconclusive (negative but low confidence) result in h-CLAT but negative high confidence result in GARD®skin. The substance met the criteria for a positive response in LLNA corresponding to KE4, T-cell proliferation. Although the LLNA result suggest that tetrakis (2-ethylbutyl) orthosilicate may meet the classification criteria for a skin sensitizer, no additional mechanistic inputs across the AOP for skin sensitization corroborate this conclusion. Taken together with negative genomic profile, i.e. GARD®skin results, and the absence of any structural alerts associated with known sensitization responses, the entirety of the evidence suggests that tetrakis (2-ethylbutyl) orthosilicate is less likely to have skin sensitization potential.
Tab. 2: Stimulation Index (SI) values and average support vector machine (SVM) decision values for case study chemicals in the LLNA and GARD™skin assay, respectively

<table>
<thead>
<tr>
<th>Dosage (%)</th>
<th>LLNA SI Mean ± S.D. (N)</th>
<th>GARD™skin SVM Mean ± S.D. (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.74 ± 0.9 (5)</td>
<td>-1.16 ± 0.4 (3) Non-sensitizer</td>
</tr>
<tr>
<td>25</td>
<td>2.54 ± 1.5 (5)</td>
<td>-1.61 ± 0.7 (3) Non-sensitizer</td>
</tr>
<tr>
<td>50</td>
<td>3.18 ± 1.0 (5)</td>
<td>-1.07 ± 0.7 (3) Non-sensitizer</td>
</tr>
<tr>
<td>100</td>
<td>3.39 ± 2.0 (5)</td>
<td>6.15 ± 1.39 (3) Sensitizer</td>
</tr>
</tbody>
</table>

a threshold for OECD TG429: LLNA 3H-methyl thymidine SI = 3
b threshold for OECD TG 442B: LLNA BrdU SI = 1.6
c excessive local irritation observed at this concentration; mean ear thickness (day 6-day 1) > 25%; irritation was not observed at 10% or 25% concentrations indicated by erythema or measures of mean ear thickness (day 6-day 1) of 11.1% and 10.5% respectively.

3.2 WOE for hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2-undecanone

Problem Formulation

Similarly to tetrakis (2-ethylbutyl) orthosilicate, the skin sensitization potential of hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2-undecanone was assessed to determine the appropriate hazard classification for regulatory registration purposes. The testing strategy was designed to meet the information requirements mandated by the regional regulatory authority, following the decision logic recommended by the regional guidance for skin sensitization hazard identification (ECHA, 2017a; United Nations, 2019). Annex VII of the Registration, Evaluation, and Authorization of Chemicals (REACH) regulation stipulates that in vitro assays should be considered prior to conducting animal-based assays (European Parliament Regulation EC No 1907/2006).

NAM data generation

Structural alerts were present for one third of the hypothesized parent structures contained within this UVCB substance, as well as multiple metabolites of the hypothesized parent structures. TIMES Skin Sensitization with autoxidation model predicted a positive skin sensitization result for one third of the hypothetical parent structures particularly those with free hydrazine functional groups. The result was not considered adequate on its own for regulatory purposes since all but one of the hypothesized structures (which produced a negative result) fell outside of the model applicability domain. Based on the nature of the substance as a UVCB (DPRA not applicable), initial data generation focused on the mechanistically based in vitro methods associated with key events 2 and 3 in the skin sensitization AOP (Figure 1); namely, a KeratinoSens™ to assess activation of keratinocytes and the h-CLAT to assess the activation of dendritic cells through measurement of markers CD86 and CD54. In the KeratinoSens™ assay hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2-undecanone was tested up to 400 ug/mL in DMSO and produced a negative sensitization response in two of three independent assays, indicating an overall negative prediction for skin sensitization. In the h-CLAT, hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2-undecanone was tested up to 67.5 ug/mL in DMSO and produced a positive sensitization response in CD54 in two of three independent assays (EC 200 = 56.99 ug/mL). CD86 was not affected by treatment with the test substance at any dose. Further analysis using the GARD™skin assay predicted hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2-undecanone to be a sensitizer based on a mean SVM decision value of 6.15 (Table 2); all three replicates produced SVM decision values above zero, the established threshold for a positive result.

Animal data generation

Because of the conflicting evidence between KE 2 and KE3, additional information for key event 4 was developed using an LLNA (OECD TG 442B). The test substance was applied to CBA/J mice at 10%, 25%, 50%, and 100% which produced stimulation indices (SI) of 1.6, 1.9, 4.0 and 3.3, respectively (Table 2). Considering an SI=1.6 as a positive indication of skin sensitization in this assay, the effective concentration (EC) at which the test substance may produce skin sensitization (i.e. EC 1.6) is calculated to be 10%. While GHS criteria does not yet exist for sub-categorization into high potency or low/moderate potency based on an SI of 1.6, the magnitude of the dose response is suggestive of moderate potency and is consistent with recent proposals for sub-categorization (Maeda and Takeyoshi, 2019).
Evidence integration
Evidence for hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylene]-, reaction products with 2-undecanone was inconclusive for protein binding (KE1) (QSAR-out of domain), negative for Keratinocyte activation (KE2) based on KeratinoSens™ and positive for dendritic cell activation (KE3) based on h-CLAT and GARD™skin. The substance met the criteria for a positive response in LLNA corresponding to KE4, T-cell proliferation.

3.3 WOE for decyl palmitate

Problem formulation
The skin sensitization potential of decyl palmitate was assessed to determine priority as a candidate substance for a product development project. Initial screening by computational methods was applied and recognizing the chemical’s low water solubility may cause negative or inconclusive results to be rejected as a basis for global regulatory hazard identification purposes, no in vitro assays were planned for use. The LLNA was utilized for hazard identification due to its wide acceptance globally for that purpose. Due to conflicting in silico and in vivo results, additional evidence from the GARD™skin assay was sought to provide additional mechanistic context to clarify known complications associated with the LLNA (e.g. marginal result).

NAM data generation

Initial in silico screening using OECD QSAR Toolbox and TIMES models suggested the chemical was unlikely to be a skin sensitizer. An assessment of physico-chemical properties and structural alerts did not give any indication of skin sensitization potential. TIMES Skin Sensitization with autodestruction model predicted a negative skin sensitization result (see supplemental data). The GARD™skin assay predicted decyl palmitate to be a non-sensitizer based on a mean SVM decision value of -1.16 (Table 2); all three replicates produced SVM decision values below zero, the established threshold for a positive result.

Animal data generation

The [3H]-thymidine-based LLNA (OECD TG 429) was used for the purposes of hazard characterization. No skin irritation was observed in a preliminary evaluation of local skin irritation in a single mouse. Decyl palmitate was tested at 10%, 25%, and 50% (% w/w) in butane, reaching a borderline stimulation index value of 3.18 (threshold SI=3) at the highest tested concentration; the calculated EC3 value was 43% (Table 2). There are various thresholds for defining potency based on EC3 values, for instance ECETOC considers an EC3 value <10% as the threshold for strong sensitizers (ECETOC, 2003) while ECHA uses a threshold of <2% to indicate strong potency (ECHA, 2017b). Under either threshold, decyl palmitate is considered a weak sensitizer.

Evidence integration

Initial in silico modeling predictions consistently predicted non-sensitizer, thus a positive LLNA result was unexpected. Individual dose groups exposed to decyl palmitate indicated a monotonically increasing proliferation, and while there was an individual animal which produced a markedly higher proliferation response than others within the highest dose group producing the positive result, it was not a statistical outlier. GARD™skin assay results predicted decyl palmitate to be a non-sensitizer, with individual replicate decision values suggesting that the genomic signature was not close to the classification profile. Given both in silico and in vitro results suggesting decyl palmitate is a non-sensitizer, and borderline (positive but with low confidence) in vivo results, the overall WOE suggests decyl palmitate is unlikely to be a skin sensitizer.

3.4 WOE for alkylated anisole

Problem Formulation
Skin sensitization potential for alkylated anisole was assessed to determine priority as a candidate substance for a product development project. In contrast to decyl palmitate, alkylated anisole is considered a UVCB, whose synthesis results in multiple isomers, and as such cannot be represented by a single molecular structure. Initial screening by computational methods was applied and given that the chemical’s low water solubility may cause negative or inconclusive results to be rejected as a basis for global regulatory hazard identification purposes, no in vitro assays were planned for use. LLNA was utilized for hazard characterization due to its wide acceptance globally for that purpose. Conflicting in silico and in vivo results lead to the generation of additional evidence in the form of GARD™skin assay to inform a WOE.

NAM Data Generation

Initial in silico analysis was conducted on a single representative isomer and was later expanded in OASIS TIMES using a range of hypothetical structures that were considered potentially present in the test material. The single representative parent structure was identified as producing a metabolite with a structural alert for the potential to form protein adducts. The metabolite reflected a stepwise metabolism wherein the anisole organic functional group was demethylated and then oxidized to a benzoquinone functional group. Based on this observation, additional model analysis was applied using a skin sensitization model in OASIS TIMES which accounts for kinetics of protein adduct generation. This model predicted all hypothesized isomers and their metabolites to be non-sensitizers (data not shown).

The GARD™skin assay predicted alkylated anisole to be a non-sensitizer based on a mean SVM decision value of -1.61 (Table 2); all three replicates produced SVM decision values below zero, the established threshold for a positive result.
Animal data generation

The [3H]-thymidine-based LLNA was used for the purposes of hazard characterization. No skin irritation was observed in a preliminary evaluation of local skin irritation in a single mouse. Alkylated anisole was tested at 25%, 50%, and 100% (% w/w) in acetone:olive oil (4:1), and exhibited a non-monotonic dose response reaching a SI of 2.04 at 25%, 5.30 at 50%, and 3.39 at 100% concentration (Table 2). One potential discordant animal was identified in the 50% exposure group, whose decays per minute response was 15,833.75 compared with the group mean (inclusive of potential outlier) 6,455. Influence of the single observation is demonstrated by the difference between the arithmetic group mean and the median: 6,455 decays per minute compared with 3,898 decays per minute, respectively.

Evidence integration

Similar to decyl palmitate, the conflicting in silico and in vivo results were unexpected. Deeper review of TIMES in silico model predictions for multiple potential isomers identified potential minority metabolites with weak sensitizing potential, though inclusion of kinetic reaction rate information these metabolites would be unlikely to result in skin sensitization. In the case of alkylated anisole, in vivo results from a single potential outlier appeared influential on the EC3 result, and consequently the classification category. GARD™skin results predicted alkylated anisole to be a non-sensitizer, with individual replicate decision values consistently below zero. Given both in silico and in vitro results suggesting alkylated anisole is a non-sensitizer, and non-monotonic dose response from in vivo results, the overall WOE suggests alkylated anisole is unlikely to be a skin sensitizer.

3.5 Comparative computational bioactivity profiling

Due to the conflicting nature of the results for case study chemicals, we evaluated whether ToxPi could provide an unbiased check on the conclusions from a WOE based on expert judgement. Evidence associated with KEs in the skin sensitization AOP were identified from the peer-reviewed literature for comparative prototypical sensitizers, non-sensitizers, and irritating non-sensitizers (Figure 3). As an example of a prototypical skin sensitizer, cinnamal consistently activated each key event in the skin sensitization AOP, where structural alerts, electrophilic chemistry and in-silico peptide binding indicate that covalent binding of skin proteins occurs (KE1), in vitro tests indicate activation of keratinocytes (KE2) via KeratinoSens™ and dendritic cells (KE3) via h-CLAT and GARD™skin (sensitizer, average decision value 11.2 (see supplemental data³), and LLNA studies demonstrate a proliferative T-cell response (KE4: EC3 ranged from 0.5-3.1%) as well as an allergic response after challenge (Adverse Outcome, AO: Guinea pig maximization test). Additional prototypical sensitizers were included in the comparison to cover a range of potencies spanning weak (aniline, eugenol, and linalool), moderate (2-MBT), and extreme (PPD) as well as a variety of skin sensitization relevant mechanisms (e.g. SN2, Michael addition, alpha-beta unsaturated ketones, quinones, thiols/disulfides).

A representative irritating non-sensitizer, nonanoic acid, showed evidence of dendritic cell activation in two distinct measures (KE3: h-CLAT and GARD™skin; weak sensitizer, average decision value of 1.15) and a weak but dose-responsive T-cell proliferative response in the LLNA (KE4) (see supplemental data). Negative results were obtained in assays assessing the remaining KEs in the skin sensitization AOP (KE1: QSAR and DPR; KE2: KeratinoSens™; AO: Guinea pig maximization assay and human maximization test (see supplemental data³). Nonanoic acid produced strong to moderate erythema and edema in all
animals during the induction phase of the guinea pig maximization assay (Johnson et al., 2011). Additional irritating non-sensitizers were included to cover a range of potencies from irritating (hexane, sodium lauryl sulfate, and xylene) to corrosive (lactic acid).

A representative negative control, propylene glycol did not activate any of the KE in the AOP: KE1, negative predictions in OASIS TIMES and DPR; KE2, negative predictions in KeratinoSens™; KE3, negative predictions in GARD™skin (average decision value of -1.37) and h-CLAT; KE4, negative result in LLNA; and AO, negative result in guinea pig maximization test (see supplemental data3). Propylene glycol did not meet the criteria for skin sensitization in any element of the AOP. Additional prototypical non-sensitizers including isopropanol, p-aminobenzoic acid, dextran, methyl salicylate and salicylic acid were included in the comparative analysis.

ToxPi ranking of biological activity based on normalized ToxPi statistics provided an additional check of the WOE conclusions for the case study chemicals (Figure 3). While unsupervised hierarchical clustering produced a clear distinction on the bookends such that higher potency prototypical sensitizers and prototypical non-sensitizers grouped as expected, the relationships were less well defined for all other chemicals in between. For example, the lower potency prototypical sensitizers (e.g., aniline, eugenol, linalool) were clustered into their own separate group with xylene, an irritating non-sensitizer. The case study test substances alkylated anisole and tetrakis(2-ethylbutyl-orthosilicate) were grouped with the prototypical non-sensitizers, while decyl palmitate was grouped with several irritating non-sensitizers in a separate arm of the dendrogram. Though the alignment was somewhat interspersed, all test chemical concluded by traditional WOE to have low potential for skin sensitization were grouped with comparative chemicals considered to be non-sensitizers. The case study substance concluded to have skin sensitization potential based on a traditional WOE approach, Hydrazinecarboximidamide, 2-[(2-hydroxyphenyl)methylen]-, reaction products with 2 undecanone, was isolated on its own arm in between a group of irritant non-sensitizers and prototypical sensitizers.

4 Discussion

Hazard classification schemes, such as UN GHS were historically developed to classify based on in vivo data such as the LLNA or guinea pig maximization assay. Considering multiple inputs collectively in a WOE approach can reduce the uncertainty associated with weakly positive results in individual animal studies through incorporation of mechanistic information. The cases presented here illustrate that consideration of the LLNA/LLNA BrdU-ELISA assay on its own would have warranted classification as skin sensitizers for all four test chemicals, however unsupervised clustering of AOP-anchored data points supports a lack of skin sensitization potential for the three test chemicals concluded to have low skin sensitization potential based on a WOE approach informed by inclusion of in vitro and in silico data.

Weak or borderline positive results in the in-vivo LLNA/LLNA BrdU-ELISA for three case study test substances (tetrakis(2-ethylbutyl) orthosilicate, decyl palmitate, and alkylated anisole) were weighted against other available data. We concluded that the three case study test substances were considered unlikely to have skin sensitizing potential based on lack of genomic response indicative of skin sensitization, e.g. GARD™skin, a lack of structural alerts and/or applicable QSAR predictions, and the unsupervised clustering conducted in ToxPi. While the absence of a structural alert is not a confirmation of absence of effect, it is difficult to gain any mechanistic insights that could place the weak positive LLNA/LLNA BrdU-ELISA results into context; however, several hypotheses were considered for the potential discordant results. The positive LLNA/LLNA BrdU-ELISA results may be related to non-specific proliferation of lymph node cells representing a false positive. As noted, none of the test chemicals had positive QSAR predictions for parent or possible metabolites by autoxidation consistent with formation of haptons, however neither tetrakis (2-ethylbutyl) orthosilicate nor alkylated anisole fit into the applicability domain for the QSAR models employed. Recent evidence suggests that the false detection rate in the LLNA is higher for test substances with a LogKow > 3.5 likely due to unspecific cell proliferation (Natsch et al., 2023). This is most clearly observed in the case of tetrakis (2-ethylbutyl) Orthosilicate, where elevated SI values were associated with excessive local irritation. Alternatively, a CD1d lipid antigen type response has been shown to stimulate T-cell proliferation, a potential possibility in the case of decyl palmitate and alkylated anisole. To the extent of our knowledge, neither decyl palmitate nor alkylated anisole nor their metabolites have been identified as CD1d ligands but are expected to be metabolized and glucuronidated. The combination of a lipid moiety attached to a sugar are the organic functional groups that are present in established CD1d ligands (McEwen-Smith et al., 2015). Finally, the resulting positive SI values described above could represent a “true” positive in the LLNA but the chemicals are still unlikely to be allergenic in humans, possibly due to physiological differences between laboratory animals and man. Certainly, the lipophilic nature of the test chemicals in this case suggests they are more likely to become trapped in the stratum corneum lipids of human skin (which is thicker than in mice) and may be unable to reach the epidermal proteins required for a reaction to induce an allergenic response (Tojo et al., 1988). The stability of the test items in the chosen vehicles has not been directly assessed, however knowledge of the chemistry of each test item would not suggest any reactive moieties.

To contextualize the weakly positive LLNA results further, additional data from skin sensitization assays were generated or identified in the public domain. A comparative class assessment of bioactivity profiles done via an unsupervised hierarchical clustering approach using ToxPi attempted to provide a non-expert judgement-based class prediction of the skin sensitization potential for each of the four case study test substances. Prototypical substances representing classes of sensitizers, irritant/non-sensitizers, and non-irritating/non-sensitizers were included for this comparison. When all AOP-based inputs were considered equally across multiple classes of skin sensitizers, the three case study test substances concluded to have low potential for skin sensitization based on an expert judgement based WOE approach were clustered together with the non-sensitizing substances, while the case study test substance concluded to have potential as a skin sensitizer by the traditional WOE approach was clustered in between the irritating/non-sensitizers and the prototypical sensitizers (Figure 3).
Our experience of this implementation of the ToxPi approach is that it has difficulty differentiating missing data from negative test data. It’s possible that the higher number of low confidence or missing data points between the case study test substances and the comparative dataset could be a contributing factor to the mixed predictions made using the ToxPi visualization. An alternative approach could be to reduce the weighting factor for assays that are more likely to produce low confidence results for difficult to test substances, such as the DPRA or h-CLAT. In this way, the low confidence results could still be considered in the overall WOE instead of excluding them altogether. In this scenario, positive outcomes would not be discounted regardless of the likelihood that the result is a true positive. Differential weighting of substances within specific endpoints is not possible in ToxPi. Other options for handling missing data would have involved imputing them with the group mean, median or mode of all the results measured for each endpoint assay. This method appears to have a larger impact when ToxPi is used for prioritization of substances which was not the focus of this work (To et al., 2018). However, missing data was not without impact in this implementation, as the case study test substance that was concluded to be positive (Hydrazinecarboximidamide, 2-[2-hydroxyphenyl]methylene]-, reaction products with 2 undecanone) was not grouped with the prototypical skin sensitizers primarily due to the lack of guinea pig data (confirmed by adding in mock guinea pig results; data not shown). Alternatively, it was not tested whether matching the dataset of the comparative chemicals to the dataset available for the case study chemicals would improve the concordance of clustering. Human data was not available for any of the case study test chemicals; in addition, we did not search for or include data for structural analogues of the test substances in this instance. However, both types of data could be easily implemented as additional data points in ToxPi given the read-across is justifiable. In our experience, using ToxPi as a comparative visualization tool could provide a useful approach to integrate lines of evidence to transparently inform class predictions with the recognition that performance is optimal when the available dataset is equivalent between the test set and the comparative known chemicals.

Recently published DAs are becoming widely recognized as first-tier testing strategies with equivalent or superior ability to predict skin sensitization potential in animals or in humans when compared to predictions from the LLNA (Kleinsteuber et al., 2018). None of the currently available DAs were applicable for our case study test substances due to the difficult to test nature of poorly soluble and UVCB substances in aqueous based test systems (OECD, 2002). As an example of this difficulty, tetrakis (2-ethylbutyl) orthosilicate was inconclusive in the DPRA due to formation of precipitate (i.e., technical limitation) and negative with low confidence in the h-CLAT due to a tendency for this assay to produce false negatives for chemicals with high LogKow (i.e., predictive limitation) (Takenouchi et al., 2013). The first published DA, the “2 out of 3” approach, requires that the DPRA and h-CLAT be used as measures of KE1 and KE3 (OECD, 2021), respectively, thus precluding consideration of this DA for tetrakis (2-ethylbutyl) orthosilicate. The alternate DA, referred to as the Integrated Testing Strategy (ITS), provides an option to rely on QSAR predictions rather than assays related to KE1 (e.g. DPRA). However, the QSAR model predictions were characterized as out of domain by the model for tetrakis (2-ethylbutyl) orthosilicate leaving only structural alerts as an in-silico option, which are not incorporated into the ITS DA. In our experience, the most likely outcome of applying DAs that rely primarily on *in silico* or *in vitro* results to poorly soluble or UVCB substances is “inconclusive” leading to the need for additional data or modified approaches. Similar issues have been noted for difficult to test cosmetic ingredients, in particular the difficulties in assessing substances considered in the overall WOE.

Careful consideration of DAs and how to integrate non-animal test methods into the current hazard classification scheme for skin sensitization is timely as these schemes are under active discussion at the international level by the United Nations Subcommittee of Experts on GHS (see UN/SCEGHS/31/INF.27/Rev.2) as well as within competent authority regulations such as the proposed updates to the CLP regulation in the EU and the New Approach Methodologies Workplan released by the United States Environmental Protection Agency (EPA) in December 2021 (USEPA, 2021: EC, 2022). As regulatory systems for hazard classification and labeling are meant to inform non-experts on a global scale, simple chemical-agnostic classification schemes are intended to drive harmonization of hazard communication. Whatever scheme is adopted should allow for flexibility to incorporate all available data and acknowledge and appropriately weight chemical-specific challenges or technical limitations of singular assays. Recent work in the new approach methodology space has focused on establishing scientific confidence in these approaches from the ground up rather than relying on comparison to traditional in vivo assays (van der Zalm et al., 2022). Recognizing the likely broad implementation of new approach methodologies in the future regulatory landscape, it should be noted that there has been little discussion on the applicability of these approaches to mixtures. Similar to complex substances, mixtures (e.g. lubricants, greases) can share the same complexity and physicochemical issues observed for our test cases and deserves further study.

Ultimately our work aims to highlight the challenges of determining skin sensitization hazard in the case of difficult to test substances with conflicting or low confidence data spanning in-vivo, in-vitro and in-silico. This initial work could have implications for future regulatory acceptance of broader assessment approaches. Each assay on its own has variability and limitations, but that does not preclude their utility to predict hazard. We demonstrate that unsupervised clustering, where the data inputs are anchored by the biological response pathway and informed by test system compatibility with the test material, can provide further support for the individual substance WOE prediction in certain cases. This combined WOE and clustering approach maximizes the contribution of all data, whether derived from *in vitro*, *in vivo* or *in silico* models, to the overall assessment of skin sensitization potential for difficult to test substances.

---

References


ECHA (2017a). Guidance on information requirements and chemical safety assessment: Chapter r.7a: Endpoint specific guidance, version 6.0. doi:10.2823/337352


Johnsson, W., Heldreth, Bart et al. (2011). Final report of the cosmetic ingredient review expert panel on the safety assessment of pelargonic acid (nonanoic acid) and nonanoate esters. *Int J Toxicol* 30, 2285-2695. doi: 10.1177/1091581811428980


USEPA - U.S. Environmental Protection Agency (2021). New approach methods work plan (v2).


**Conflict of interest**

The authors are currently or formerly employed by ExxonMobil Biomedical Sciences, Inc., a separately incorporated but wholly owned affiliate of ExxonMobil Corporation, a company with commercial interest in the case study substances included in the publication. Some authors may also own stock in publicly traded companies with a commercial interest in the case study chemicals. The manuscript was written as part of their normal employment and was the sole responsibility of the authors. No external funding was obtained for manuscript preparation.

**Acknowledgements**

Supplemental and raw data files are available on the Center for Open Science: https://osf.io/89ez4/

**Data availability**

The authors would like to thank Christine Palermo, PhD. of ExxonMobil Biomedical Sciences Inc., and Tim Lindberg of SenzaGen for helpful discussion and input during development of the manuscript.